

Testing for Justice

by Edward G. Rozycki

There is one story that, over the past seventy years, increasing numbers of school people in the United States have come to tell. The story expresses widely shared aspirations and deeply felt concerns. Let us call this story the “Testing for Justice Rationale.” It goes something like this:

For schools to meet the needs of some children and not others is unfair. Justice therefore dictates that we meet the needs of all children. How do we determine those needs? By comparing what children can do with what they can learn to do. Any discrepancy between achievement and potential indicates a need. How do we determine such questions as achievement and potential? Through adequate testing.¹

That rationale, though it supports many well-intentioned attempts at upgrading American schools, is replete with questionable assumptions seldom examined after repeated tries at improving schooling practice have failed.

Educational testing has long been noted to affect the lives of not only students but educators themselves.² Thus, an understanding of testing and the assumptions on which it is based is indispensable to intelligent schooling practice. Tests can be critiqued not only for their technical efficiency but also on whether they are fair and whether the very process of testing is little more than an exercise of political power. The Testing for Justice Rationale burdens testing with determining not only need but, ultimately, justice as well.

Why Have Tests?

Modern schooling, which processes large numbers of students, seems inconceivable without testing. That is because it is so convenient for sorting students. It can stand in for a long and involved set of social interactions with master teachers—more typical of an apprenticeship

and more common before schools grew to present-day sizes and pursued a philosophy of productive efficiency.

Why do teachers give tests? For several reasons, among them: a) to support the authority of the teacher's judgment about acquired learnings, and b) to substitute for an infeasibly broad examination of student ability. This convenience is so important in the mass processing of today's schools that learnings not susceptible to easy examination, e.g., with paper and pencil, find it hard to gain status in a curriculum. Goodson comments,

For the groups and associations promoting themselves as school subjects, and irresistibly drawn to claiming academic status, a central criterion has been whether the subject's content could be tested by a written examination for an able clientele.³

In testing, however, we make many crucial assumptions about means, ends, and the causal connections among them. Achievement tests, for example, are not in and of themselves the point of instruction; otherwise, we would teach, not merely *to the tests*, but the very tests themselves. Nor is mere participation in course work thought sufficient to make testing unnecessary. Rather, the ends sought in achievement tests are certain important residues of the instructional process.

Calling something a test assumes a strong consensus on what its results indicate. But for constructs as vague and controversial as human abilities, upon which a judgment of educational need might be based, interesting things happen. On one hand, tests may stand in for controversial and pluralistic conceptions of human ability. Intelligence, for example, becomes what IQ tests measure. On the other hand, the concept of, say, intelligence itself becomes a focus of controversy.⁴

What Makes a Test a Test?

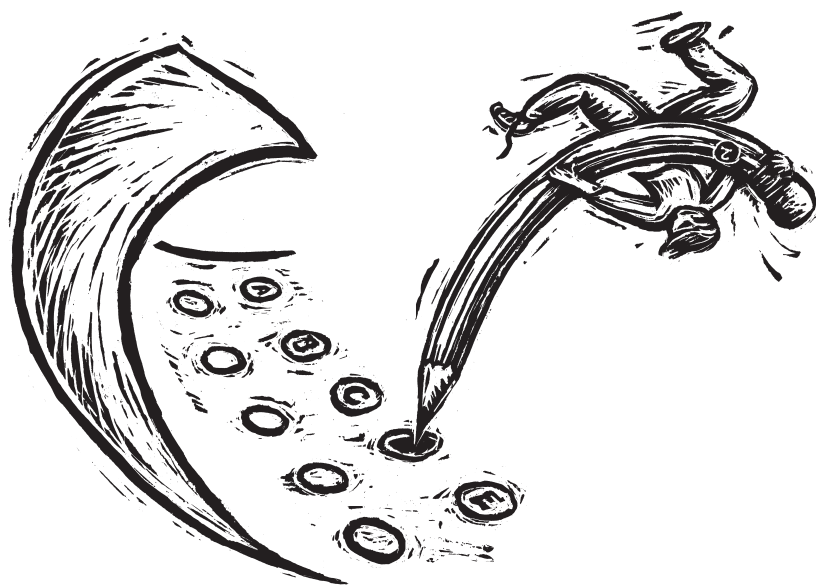
From the student point of view every test is a task.⁵ But not every task is a test, even if it looks like one. What conditions must a task satisfy to constitute a true test? It is a question of great practicality. State governments base school district funding on efficiency, itself determined by tests that state departments of education impose on the districts. But what will make the procedure anything more than a charade?

To avoid overlooking assumptions built into our conception of testing, let's substitute a different concept, *rank-task*, for tests. A rank-task is a type of activity for which some outcomes can be ranked: better, the same, or worse. Think of a rank-task as any procedure that assigns a number. It can be interpreted as a rank to compare that person to others involved with the procedure. Cinderella's prince, looking to fit the glass slipper, would be undertaking such a rank-task. Some feet are too small;

others, too large; only Cinderella's, just right. Trying to sort football players by the numbers on their jerseys is not a rank-task, though, because there is generally no significance to the comparison of any two numbers other than indicating a different wearer.

Tests are, at the minimum, rank-tasks. They can be performed with more or less skill. But the skill demonstrated may not be what we wish to measure. For instance, students take SAT-preparation courses to learn test-taking skills, not the information the tests are designed to measure. Often test-taking skills can be as critical to earning a good score as actual knowledge of the material the test covers. For some years, for example, the Princeton Review's basic test-taking materials and training have evidently raised SAT scores significantly.⁶ The SATs are intended to measure scholastic aptitude, but the effectiveness of the Princeton Review materials suggests that the SATs are also measuring something else—namely, the ability to take standardized tests of this type.

That observation illustrates the practical nature of our seemingly theoretical observations about testing. Among the readers of this article there certainly are individuals who did not receive a scholarship, or who were not accepted to the college or university of their choice, because of the scores they received on the SAT. And there is a fair chance that the reason for those scores was lack of, not scholastic aptitude, but certain test-taking skills.



Tests are also taken to be indicators. As such they must meet certain conditions. Any well-designed rank-task must be able to vary consistently upon reapplication, and the variation must be understood to make a difference. Test makers call that trait consistency, or internal validity. Tests must also indicate something other than themselves: that is called “external,” or construct, validity.⁷

Usually out of the hands of professional test makers is a fourth condition, trustworthiness: we must be able to believe the results were not manipulated for special purposes. That is usually a matter of test security, a matter with which many schools not infrequently deal in cavalier fashion.⁸

The important point, especially from the test taker’s point of view, is that every test is a task that can be performed, independently of such technical considerations as externality and trustworthiness, with greater or lesser skill. For example, a student may learn to take multiple-choice exams efficiently even if those exams test nothing recognizable as subject matter; yet a student who knows a great deal about some subject may falter at demonstrating that knowledge on the test prepared for it.⁹

If a rank-task is a test, then the goals of the testing control (determine) the kinds of test tasks we present to the student. Those tasks in turn control the knowledge the student will have to bring to support the test task. The connections between student knowledge and the test outcomes used to evaluate it are mediated by the task itself. Whether an increase in test scores indicates an increase in student knowledge or an increase in test-taking skills may depend on such mediation.

From Consensus, through Testing, to Justice

Let’s reiterate an important point: we, as interested parties, must agree upon some way of determining student knowledge independent of the test; otherwise the test becomes problematic. Lacking such consensus on the test, evaluations of potential or achievement are questionable. So then is the determination of need and consequently fairness.

Thus, in a very real way, problems of consensus are what bear ultimately, via testing, upon perceptions of fairness in schooling. We can lay the argument out as follows:

- a) Consensus, among interested parties, will affect which ideas of potential (e.g., native ability, capacity, competence) can be used for testing in the school.
- b) Consensus will affect which ideas of achievement (e.g., skills acquired or developed) can be used for testing in the school.

We then bring in the connections given by the Testing for Justice Rationale:

- c) The difference between potential and achievement measures need.
- d) The difference in treatment of need measures justice.

The most immediately practical version of this argument, which we will call the Status Quo Argument, is this:

There is a consensus in our community that Group A and Group B differ in potential. We observe that they differ in achievement. Since their achievement merely reflects their potential, there is no disparity in educational need. Therefore, our present treatment of Groups A and B, although they may look different, is not unjust.

The Status Quo Argument is theoretically sound, despite the fact that it has been pressed into the service of racism and class bias.¹⁰ The moral issue is how its supporting consensus arises. It is around such claims of consensus that many of the controversies about schooling cluster. (Consider, for example, the widely accepted assumption that so-called “gifted students” have no need for special educational treatments.)

Objectivity and Need

One assumption of much discussion about schooling practice is that testing offers an objective decision-making procedure that avoids problems of values and consensus. But is that so?

Test data seem so impartial, so objective. But what can numbers alone tell us? Imagine three groups of students, A, B, and C, who each receive a rank-task: Rank-Task 1, Rank-Task 2, and Rank-Task 3.

Suppose chart 1 gives us the following results—assuming the group means to be calculable.

Chart 1

	Rank-Task 1	Rank-Task 2	Rank-Task 3
Group 1	95	95	60
Group 2	50	50	60
Group 3	15	15	60

Even if we can also assume the significance of intergroup differences for each test and the absence of cheating, what are we to make of the differences in these scores? Are they any guide to practical decision-making?

It all depends. Our first question should be “What are these tests supposed to indicate?” Unless we believe they indicate something, they are still merely rank-tasks. And test results that are important to making equitable schooling decisions must deal with what Thomas Green has called *educationally relevant attributes*.¹¹ An attribute is educationally relevant in Green’s terms if it would be fair to distribute schooling benefits based on that attribute. If we believed it fair, for example, that males receive more diplomas than females just because they are males, then sex would be an educationally relevant attribute.

In America, unlike some other cultures, sex is by law not educationally relevant in public schools. Let us imagine a society so fixated on gender stereotypes that psychological distinctions override the physiological. To the extent that a female were seen as a “tomboy,” she would receive preference with “real men” over other females. “Girly men” would be devalued. In such a society, the test that decided who enjoys the privileges of gender prejudice would be called the “Degree of Masculinity Test.”

In chart 1, suppose Test 1 indicated something like “degree of masculinity” (DMT). If Test 2 indicated the percentage of high school graduates in the group, we in the United States would find that it indicated an unjust situation, because we reject gender as educationally relevant. But if instead Test 3 stood for the percentage of high school graduates, it would be taken, on the same assumption of the irrelevance of gender, to indicate equitable schooling practice.

More Educationally Relevant Attributes

Chart 2 (see next page) shows attributes in terms of which people might be grouped compared with different kinds of schooling benefits. In each block the words *just* or *unjust* indicate whether there is a general consensus in the United States that any schooling benefits distributed by the indicated kinds of attribute are considered just. Question marks indicate controversial practices.¹²

The chart indicates that in different situations an attribute may be educationally relevant, or it may not. Consider the case of sex-based grouping for varsity sports. Sex is generally not considered a relevant attribute so far as any educational benefit is concerned. Distributing high school diplomas based on sex, for example, is unjust. But participation in varsity sports is another matter. There is sometimes controversy about allowing women to play football, particularly in public high schools. Our chart indicates that with a question mark. (Imagine how chart 2 would look if it reflected common opinions in the United States circa 1800.)

Chart 2

ATTRIBUTE	BENEFITS					
	High school diplomas	Access to further schooling	Knowledge per se	Playing varsity sports	Nurturance	Special Programs
SEX	unjust	unjust	unjust	?	unjust	just
RACE	unjust	unjust	unjust	just	unjust	just
HEIGHT	unjust	unjust	unjust	just	unjust	just
ABILITY	just	just	just	?	unjust	just
EFFORT	?	?	just	just	unjust	just
CHOICE	just	just	just	just	?	just
NEED	unjust	unjust	unjust	unjust	?	just
WEALTH	?	unjust	just	?	unjust	just
DISABILITY	just	just	?	just	unjust	just
POTENTIAL	just	just	just	just	unjust	just
ACHIEVEMENT	just	just	just	just	unjust	just

Choice is an important and controversial attribute in our culture. It is not generally considered unjust if adults who decline to participate in certain programs, for example, fail to benefit from those programs. Lack of participation by children or mentally incompetent persons is often taken as a sign of immaturity or incompetence. Truancy is an example. Significantly, the lack of consequent benefits in truancy is still often argued as unjust, despite the insinuation that coercion may be justified. (This sense of injustice no doubt supports compulsory-schooling statutes.)

Other controversial practices suggested by the chart are:

- a. allowing students to play varsity sports on the basis of choice (interest) rather than ability (a long-established practice at Swarthmore College);
- b. social promotion—promoting students on effort rather than knowledge;
- c. providing nurturance, a scarce resource, on need rather than traditional practices of sharing per capita (i.e., “special education”);
- d. providing diplomas and sports participation based on wealth. (That is an important service of some kinds of private schooling.)

Needs and Consensus

Embedded in the Testing for Justice Rationale is an interesting equation:

(Ability) - (Achievement) = (Need)

Read this as “Ability minus achievement equals need” or “The measure of need is indicated by the difference between ability and achievement.” The equation often sorts students into three types: underachievers, normal achievers, and overachievers. Chart 3 shows several hypothetical scores for tests of ability and achievement. Using the equation given above, need is calculated. Based on need, students are typed as overachievers, normal achievers, and underachievers.

Chart 3

	Ability	Achievement	Need	Type
Group A	50	95	-45	Over-achiever
Group B	50	50	0	Normal achiever
Group C	50	15	35	Under-achiever

So it is argued that underachievers have greater educational needs—and the numbers make it seem quite objective.

Vague formulas like the one above, which can be discerned in the rationales offered, guide a surprising amount of daily school practice.¹³ They express not only accepted generalizations from practice but also conceptions of human nature. Their usefulness is not that they provide exact measures of important pedagogical constructs, but that they can so readily guide practice. But do they really identify needs? It depends on what we mean by needs.

In schooling, needs have long been treated as independent of consensus. But underlying much discussion of needs is the assumption that something should be desired. Someone who refers to a need is often urging action to address it while begging the crucial question of why we should address it.

We can distinguish between two conceptions of need: a conditional concept and an approval concept. A clear picture of the distinction can be obtained by comparing the following situations:

Situation 1: Johnny asks you to borrow a permanent marker. "I need it to write graffiti on the boys' room wall," he explains.

Situation 2: Mark tells you, "I need a permanent marker to complete my school art project."

We would deny that Johnny needs a permanent marker but concede that Mark needs one. Why? Because we do not approve of graffiti, but we value art projects. If our values were different, our assessment of needs would be different.

The conditional concept of need says merely: some item X is necessary to bring about some other item Y. The permanent marker stands in this relation to covering the wall with graffiti as it does to doing the art project. In the conditional sense, both Johnny and Mark have needs, just as cars need fuel or terrorists need explosives. A conditional need indicates, at most, a lack. But lacks do not necessarily beg for remediation.

Talking about needs in schooling transforms an objective, take-it-or-leave-it conditional need into a need that elicits our support without careful consideration. The common technique is to show a lack of some kind and then to treat that lack as synonymous with an approval concept of need. A typical instance goes something like this:

Researchers working for one or another special-interest group announce with alarm that there is a great need to emphasize classical antiquity in the high school curriculum because 97 percent of five thousand high school seniors surveyed nationwide

could not identify Achilles, the Acropolis, Adonis, Aeneas, the *Aeneid*, and several dozen similar items.

If the research has been done properly, it does demonstrate that high school seniors lack knowledge of antiquity, but it does *not* demonstrate that we should do anything about it. That is an entirely different matter.

We are not disparaging needs slogans, merely reiterating the point that they assume and obscure issues of value and consensus. If, for example, people agreed on the value of “self-fulfillment” and what it means, then they would approve of what they believe is a causal or logical necessity to achieve self-fulfillment. An even more important consideration, though, is this: people may appear unmoved by appeals to needs not by heartlessness but by different values or beliefs in what is causally or logically related.

Examining the Rationale

Suppose candidates for school positions, teachers, principals, or superintendents were asked to comment on the Testing for Justice Rationale during their employment interviews. I would wager that were they to disavow or deny it, they would be denied employment (more likely, surreptitiously denied—moved to the bottom of the list—since we like to flatter ourselves that we are open to diversity in philosophy as well as race, religion, ethnicity, disability, or sexual preference—and lawsuits are expensive).

Too many schools, though, adopt such slogans as “All children can learn” or “We are dedicated to excellence”—and mandate that their staffs accept them. That leaves little wiggle room for those who find that the Testing for Justice Rationale presumes a near-blasphemous omnipotence.

Actually, if we analyze the Testing for Justice Rationale, we can see just where to distinguish issues of value versus issues of power. By doing so we may achieve consensus on important values without necessitating commitment to the possibly counterfactual optimism expressed in the Rationale. To revisit the Rationale:

For schools to meet the needs of some children and not others is unfair. Justice therefore dictates that we meet the needs of all children. How do we determine those needs? By comparing what children can do with what they can learn to do. Any discrepancy between achievement and potential indicates a need. How do we determine such questions as achievement and potential? Through adequate testing.

Is it really unfair for schools to meet the needs of some children and not others? Does the concept of readiness—so important to reading

teachers—not indicate our recognition that schools can meet the needs of the “ready” children better than others? And even if there is unfairness here, must it be schools that are responsible for addressing it? Does Justice dictate that? Or is it that other institutions in our society have foisted that off on the schools?

Ought we to accept responsibilities beyond the reasonable scope of our knowledge? In the long run, less blather about “determining potential” and more humility might enhance our professional repute to a greater extent than our posing as modern shamans for all things academic.

And if we are to accept such responsibilities, can we expect to be given reasonable resources to support our efforts? So far as funding is concerned, special education has been reneged on since its inception. Do we really expect a more generous flow from the public coffers in the future?

Testing is a side issue. Tests are constructed after most of the important issues—values, ethics, politics—that impinge upon schooling have been settled. That is why private and parochial schools are seldom consumed with the furor, the enthusiasm, and the dismay that testing brings to public education.

We may well continue to concern ourselves with the inequities we perceive in our society. We may well continue to pursue a dream of gleaming alabaster cities undimmed by human tears. If so, we might do better to look elsewhere than to public education to address our aspirations.

Notes

1. Cf. Nathan Glazer, “IQ on Trial,” *Commentary* (June 1981): 51–59, for a critique of *Larry P. v. Wilson Riles*, in which the Testing for Justice Rationale plays a prominent part.

2. Lynn Olson, “Tests Found Barring Thousands of Minority Teacher Candidates,” *Education Week* 8(12) (November 13, 1988): 1.

3. Ivor Goodson, *School Subjects and Curriculum Change* (Philadelphia: Falmer Press, 1987), 25.

4. For longstanding controversy on the nature of IQ, compare Arthur R. Jensen, “How Much Can We Boost I.Q. and Scholastic Achievement?” *Harvard Educational Review* (Winter 1969), with J. P. Guilford, *The Nature of Human Intelligence* (New York: McGraw-Hill, 1967). See also Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligence* (New York: Basic Books, 1993).

5. Cf. Gilbert Sax, *Principles of Educational and Psychological Measurement and Evaluation*, 4th ed. (Belmont, Calif.: Wadsworth, 1997), 15: “A test is a task or series of tasks used to obtain systematic observations presumed to be representative of educational or psychological traits or attributes.”

6. Adam Robinson and John Katzman, “The Princeton Review,” in *Cracking the System: The SAT* (New York: Villard Books, 1986). See also the authors’ Web site at <<http://www.princetonreview.com/law/testprep/testprep.asp?TPRPAGE=17&TYPE=LSAT-PREPARE&adcode=23089>>.

7. For example, see Jum C. Nunnally, *Psychometric Theory* (New York: McGraw-Hill, 1967), 100–102.

8. See the NASSP's online article "High Stakes Cheating" at <<http://www.naesp.org/ContentLoad.do?contentId=151>>. Other subversive factors, such as cramming, are problematic. The Internet offers a surfeit of advice on how to cram effectively as well as admonitions that it will not work. See, for example, Eastern Illinois University's Learning Assistance Center's article "Final Exams and Cramming," at <<http://www.eiu.edu/~lrnasst/finals.htm>>.

9. See Linda M. Lance, "The Effects of Teaching Test-Taking Strategies to High School Special Education Students on Achievement Scores on the New Jersey High School Proficiency Assessment" (Diss., Widener University, 2004). Its abstract is at <<http://muse.widener.edu/~egrozyck/Dissertations/LanceR.html>>.

10. An inversion of this argument used to support giving more resources to those identified as "special" argues that such a differential is not unfair to "normal" children.

11. See Thomas F. Green, *Predicting the Behavior of the Educational System* (Syracuse, N.Y.: Syracuse University Press, 1980), 49–52.

12. These results are meant merely to be illustrative for the purposes of the exposition.

13. Other slogan formulas commonly encountered are *potential + learning = ability*, *ability + motivation = performance*, *achievement = sum(performances)* and *potential ≥ achievement*.

*Edward G. Rozycki, Ed.D., a contributing editor for **educational HORIZONS**, is an associate professor of education at Widener University, Widener, Pennsylvania.*

Publishing in educational HORIZONS®

educational HORIZONS requests that potential contributors or guest editors consider the topics in the adjoining space or suggest related themes or articles for upcoming issues of the journal.

Book reviewing:

Individuals interested in submitting book reviews (including more substantial book review essays that would review relevant scholarship on the topic) are encouraged to query by first-class letter or e-mail. Proposals, which can be independent of our issue themes, should specify recent book releases that will interest our readership of teachers and teacher educators.

For more information, contact:

educational HORIZONS®
P.O. Box 6626
Bloomington, IN 47407-6626
publications@pilambda.org

- Where Is Public Schooling Headed?
- Teacher Education: What Must Be Done?
- How Pertinent Is Education Research?
- The Great Grade Inflation, Its Causes, and Its Effects
- America's School Funding Fiasco: The Evils of Inequity
- Are Textbooks Obsolete?
- Silk Purses from Sow's Ears: Are the Broad Masses Truly Educable?
- The Classroom Impact of Multiple Intelligences Doctrine
- A Report Card for No Child Left Behind
- Are We All Constructivists Now?
- The Reading Wars: Where Are We Now?